DOCUMENT RESUME

ED 473 529 TM 034 737

AUTHOR	van der Linden, Wim J.
TITLE	Some Alternatives to Sympson-Hetter Item-Exposure Control in Computerized Adaptive Testing. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
SPONS AGENCY	Law School Admission Council, Princeton, NJ.
REPORT NO	RR-02-02
PUB DATE	2002-00-00
NOTE	36p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. E-mail: Fox@edte.utwente.nl.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; College Entrance Examinations; *Computer Assisted Testing; Law Schools; Probability; Test Construction; *Test Items
IDENTIFIERS	*Item Exposure (Tests); *Law School Admission Test

ABSTRACT

The Sympson and Hetter (SH; J. Sympson and R. Hetter; 1985; 1997) method is a method of probabilistic item exposure control in computerized adaptive testing. Setting its control parameters to admissible values requires an iterative process of computer simulations that has been found to be time consuming, particularly if the parameters have to be set conditional on a realistic set of values for the examinees' ability parameter. Formal properties of the method are identified that help us explain why this iterative process can be slow and does not guarantee admissibility. In addition, some alternatives to the SH method are introduced. The behavior of these alternatives was estimated for an adaptive test from an item pool from the Law School Admission Test. Two of the alternatives showed attractive behavior and converged smoothly to admissibility for all items in a relatively small number of iteration steps. (Contains 4 figures and 14 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 473 529

Some alternatives to Sympson-Hetter Item-Exposure Control in Computerized Adaptive Testing

Research Report 02-02

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Nelissen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Wim J. van der Linden

TM034737

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of Educational Measurement and Data Analysis

Some Alternatives to Sympson-Hetter Item-Exposure Control in Computerized Adaptive Testing

Wim J. van der Linden

This study received funding from the Law School Admissions Council (LSAC). The opinions and conclusions contained in this paper are those of the other and do not necessarily reflect the policy and position of LSAC. The paper was completed while the autor was a Fellow at the Center of Advanced Study in the Behavioral Sciences, Stanford, CA. The author is indebted to the Spencer Foundation for a grant awarded to the Center sepport his Fellowship. The computational assistance of Wim. M.M. Tielen is gratefully acknowledged.
Requests for reprints should be sent to W.J. van der Linden, Department of Educational Measurements and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, THE NETHERLANDS. Email: w.j.vanderlingen@edte.utwente.nl

Abstract

The Sympson and Hetter (1985; 1997) method is a method of probabilistic item-exposure control in computerized adaptive testing. Setting its control parameters to admissible values requires an iterative process of computer simulations that has been found to be time consuming, particularly if the parameters have to be set conditional on a realistic set of values for the examinees' ability parameter. Formal properties of the method are identified that help us explain why this iterative process can be slow and does not guarantee admissibility. In addition, some alternatives to the SH method are introduced. The behavior of these alternatives was estimated for an adaptive test from an item pool from the Law School Admission Test (LSAT). Two of the alternatives showed attractive behavior and converged smoothly to admissibility for all items in a relatively small number of iteration steps.

Key words: computerized adaptive testing; item-exposure control; iterative adjustment of control parameters; Sympson-Hetter method.

Some Alternatives to Sympson-Hetter Item-Exposure Control

in Computerized Adaptive Testing

The Sympson-Hetter (SH) (1985; see also Hetter & Sympson, 1997) method is the most popular method of item-exposure control in computerized adaptive testing (CAT). The method is based on a probabilistic experiment that is used to determine if an item that is selected should be administered. The conditional probabilities of item administration given selection of the item are the control parameters used to restrict the item-exposure rates. The values of these parameters have to be set through an iterative adjustment process in which at each step the effects of the previous adjustments are estimated using computer simulations of adaptive test administrations.

In practical settings, the use of the SH method has been found to be time consuming. Typically, the method is applied to control the exposure rates of the items in the pool conditional on 10-12 possible values for the examinees' ability parameters, θ. The number of iterated CAT simulations required to find admissable values for the control parameters for one θ value is generally of the same order. It is therefore not unusual to have to run some 100-150 computer simulations before the SH method can be used operationally in a CAT program. If the item pool is changed, for example, because some of its items appear to be flawed or have been the victim of a security breach, the process has to start all over again (Chang & Harris, 2002).

When using the SH method, it is regularly found that at several iteration steps some of the exposure rates of overexposed items increase rather than decrease. Also, sometimes exposure rates that have been brought below a target value jump back to larger values at later steps. Further, no matter the number of iterations steps, occasionally it appears impossible to get all exposure rates below a target value that nevertheless seems reasonable. Because of this behavior of the SH method, it is necessary to eyeball the item-exposure rates in the iterative process and use our personal judgment to decide when to stop.

It is the purpose of this paper to present some alternative methods of item-exposure control that are all based on the same idea of adjustment of control parameters through an

iterative process of computer simulations. These methods were derived using an analysis of the formal properties of the adjustment rule in the SH method. Two of the methods showed particularly attractive behavior and converged directly to admissibility for all items in a relatively small number of steps.

Computerized Adaptive Testing

Let $i = 1, ..., I$ denote the items in the pool and $k = 1, ..., n$ the items in a specific adaptive test. In the empirical examples later in this paper, the items in the pool fitted the 3-parameter logistic (3PL) model. According to this model, the probability of an examinee with ability level $\theta \in (-\infty, \infty)$ on item i is equal to:

$$p_i(\theta) \equiv \Pr(U_i = 1 \mid \theta) \equiv c_i + (1 - c_i)\frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \tag{1}$$

where $b_i \in (-\infty, \infty)$, $a_i \in [o, \infty)$ and $c_i \in [0, 1]$ are parameters that represent the difficulty, discriminating power, and the guessing probability on item i, respectively (Birnbaum, 1968).

Prior to the selection of item k the estimator of the examinee's value for the ability parameter is updated using his/her responses on the previous $k - 1$ items. The updated estimate is denoted as $\widehat{\theta}_{k-1}$. In the empirical examples below, $\widehat{\theta}_{k-1}$ was the mean of the posterior distribution of θ (EAP estimator).

The usual criterion to select items in CAT is the maximum-information criterion. Let $I_i(\theta)$ denote Fisher's information measure at θ, i_k the item in the pool that is administered as the kth item in the test, and R_k the subset of items from the pool that is available to chose item i_k. The maximum-information criterion is given by

$$i_k \equiv \arg\max_j \left\{ I_{j_k}(\widehat{\theta}_{k-1}) : j \in R_k \right\}, \tag{2}$$

where j is a dummy variable indicating the items in R_k. For more details on the EAP estimator and maximum-information criterion as well as on alternative estimators and

criteria for use in CAT, see van der Linden and Pashley (2000) and Wainer (2000, chap. 4-5).

Because the criterion in (2) involves maximization over a set of discrete items and the item information functions are real valued and smooth, item selection in CAT typically favors the same item over small intervals of θ values. In addition, the number of intervals is usually much smaller than the size of the pool, and CAT algorithms therefore have a tendency of overexposing a small number of items in the pool and ignoring the remaining items, often leading to a distributions of exposure rates that can be described by Zipf's law (Wainer, 2000). The SH method has been introduced to adjust the exposure rates of items with a tendency to overexposure.

One of the consequences of imposing control on items with tendencies to overexposure is a possible reduction of the information on the examinees' ability levels in the test. The presence of control means that the algorithm cannot always pick the item with the maximum value for the selection criterion in (MI). In such cases, the amount of information lost need not be large if enough high-quality items are available in the pool and the algorithm can settle for items only slightly less informative than the item with maximum information. However, further exploration of this topic, which should not only take into account the composition of the item pool but also on the presence of the other constraints on the item selection, e.g., on test content, is beyond the scope of this paper.

Sympson-Hetter Method

Let S_i denote the event of selecting item i, and A_i the event of administering item i. For $i = 1, ..., I$, it holds that

$$A_i \subset S_i, \tag{3}$$

Therefore,

$$P(A_i) \leq P(S_i). \tag{4}$$

and

$$P(A_i) = P(A_i, S_i) = P(A_i \mid S_i)P(S_i) \tag{5}$$

The goal of the SH method is to get the item-exposure rates, $P(A_i)$, below an upper bound r_{max}, that is, to get

$$P(A_i) \leq r_{max}, \text{ for } i = 1, ..., I \tag{6}$$

It is possible to introduce different upper bounds for different items in (6), but this option will not be explored here. Several formal properties of item exposure rates in CAT as well as of the way they are controlled in the SH method are summarized in the Appendix.

The value of r_{max} should be set at a realistic level. The lower bound in (A2) in the Appendix requires that r_{max} should never be set smaller than nI^{-1}. In real-life CAT, item pools are typically 7-10 times the length of the adaptive test. The corresponding range of nI^{-1} is .10-.14. This range fits well with the target values for high-stakes tests which are typically set in the range of .20-.30 for the period of time the item pool is operational.

The probability of item selection, $P(S_i)$, depends on a variety of factors including (1) the IRT model; (2) the CAT algorithm; (3) the choice of the initial item; (4) the ability estimator; (5) the composition of the item pool; and (6) the ability distribution in the population of examinees. Once a test is operational, these factors are fixed both by the design of the CAT and the ability distribution of the examinees in the population. However, during test administration it is possible to manipulate the control parameters $P(A_i \mid S_i)$ such that the goal in (6) is attained. We will refer to a set of values for the control parameters that attain the goal in (6) as *admissible values*.

Note that we can always realize admissibility by choosing low values for $P(A_i \mid S_i)$ for all items. However, implicit in (5) is the idea that $P(A_i)$ should be below r_{max} but that for items with a tendency of overexposure $P(A_i)$ should not be brought too far below this bound. These items generally have good measurement properties, and it would be a

pity to loose them entirely. This requirement is hard to formalize but will be assumed to hold in the remainder of this paper.

It may even seem ideal to have exposure rates slightly below the target not only for the items with a tendency to overexposure but for all items in the pool. In this case, we would still protect the item pool from compromise but at the same time have maximum return on the investments in item writing and calibration. However, the upper bound in (A3) shows that this goal is unrealistic. As discussed earlier, this bound typically is a low number in the range of .10-.14. Because a small subset of items tends to have exposure rates close to a target value in the range of .20-.30, it even follows from (A1) that a substantial numbers of items in CAT pools must have negligible exposure rates and the actual minimum exposure rate can be expected to be equal to zero.

Adjustment of Control Parameters in the SH Method

The control parameters in the SH method can not be solved analytically for a set of admissible values. Sympson and Hetter introduced an iterative rule for adjusting the values of the control parameters through a series of CAT simulations that have to be continued until admissibility is obtained. The simulations are necessary to estimate the changes in the probabilities $P(S_i)$ and $P(A_i)$ after each adjustment. The iterative rule can be defined as follows. Let t denote the iteration steps, $P^{(t)}(A_i \mid S_i)$ the value of the control parameter for item i at Step t, and $P^{(t)}(S_i)$ and $P^{(t)}(A_i)$ the probabilities of selecting and administering item i at Step t. If the simulation at Step t is completed, $P^{(t)}(S_i)$ and $P^{(t)}(A_i)$ are estimated, and for items for which the estimates of $P^{(t)}(A_i)$ do not meet (6), the values of the control parameters are adjusted. The SH adjustment rule is as follows:

$$P^{(t+1)}(A_i \mid S_i) := \begin{cases} 1 & \text{if } P^{(t)}(S_i) \leq r_{\max}, \\ r_{\max}/P^{(t)}(S_i) & \text{if } P^{(t)}(S_i) > r_{\max}, \end{cases} \tag{7}$$

$i = 1, ..., I.$

This rule is based on the following argument: If at Step t item i was selected by the CAT algorithm with a probability smaller than $r_{\max}$, the inequality in (4) implies that the

item also had an exposure rate smaller than r_{max} and does not need any control. Hence, $P^{(t+1)}(A_i \mid S_i)$ can be set equal to 1. However, if item i was selected with a probability larger than r_{max}, its control parameter should have been set such that $P(A_i) = r_{max}$. From (5) it follows that this requirement would have been realized if $P^{(t)}(A_i \mid S_i)$ had been equal to $r_{max}/P^{(t)}(S_i)$. Hence, $P^{(t+1)}(A_i \mid S_i)$ is set at this value. A more concise representation of this argument is given in (A6), where it is also indicated that the argument rests on the (invalid) assumption that the equality in (5) holds between probabilities in different iteration steps.

Behavior of Exposure Rates during Adjustment

As already noted, the adjustment process for the control parameters in the SH method is generally time consuming, can show unexpected behavior of exposure rates for some items at some steps, and may have difficulty converging to admissibility at all. The constraint on the probabilities of item selection in the pool in (A5) explains much of this behavior of the SH method. A major implication from (A5) is that if, at Step t, the only adjustments of the exposure control parameters are negative, we always have positive effects on the probabilities of item selection for some of the items at Step $t + 1$. On the other hand, if the adjustments at Step t are positive, we always have negative effects on some of these probabilities at Step $t + 1$. More formally, it thus holds that if

$$P^{(t+1)}(A_i \mid S_i) < P^{(t)}(A_i \mid S_i)$$

are the only adjustments for some nonempty set $i \in U$, then

$$P^{(t+1)}(S_i) > P^{(t)}(S_i), \tag{8}$$

for some nonempty set of items $i \in V$, where the implication remains true if the inequalities are reversed.

One consequence of (8) is that the effect of a negative adjustment of the control parameter for an item at Step t may be an increase of the probability of selection, and

hence of the exposure rate, for the *same* item at Step $t + 1$. The SH method then moves the exposure rate in the wrong direction and does not convergence for this item. Also, (8) implies that the effect of a negative adjustment of the control parameter for one item can be an increase of the probability of selection and hence a return of the exposure rate of another item to a value larger than r_{max}. In principle, it is thus possible to have two items with exposure rates jumping back and forth between values in the intervals $[0, r_{max}]$ and $(r_{max}, 1]$. Again, the SH method then has difficulty converging to admissibility. Finally, (8) shows that if the control parameters of several items are adjusted at the same time and some of the adjustments are negative and others positive, the effects on the probabilities of selection, and thus on the exposure rates, at the next step neutralize each other. One possible result is that the SH method may move in the right direction but progresses only slowly.

Alternatives to the SH Method

We now view the SH method purely as an algorithm for producing admissible values for the exposure control parameters, and would like this algorithm to have the following properties: (1) adjustments of the control parameters should be only for items with exposure rates above the target; (2) the adjustments should be effective, that is, in one step; and (3) once the control parameters have been adjusted the exposure rates should not be allowed to jump back to values larger than the target at later steps.

We will try to get as closely as possible to these properties by defining alternative algorithms that have combinations of levels of the following factors: (1) positive adjustments or no adjustments if the items have exposure rates below the target; (2) using $P^{(t)}(S_i)$ or $P^{(t)}(A_i)$ in the definition of the subsets of items that need adjustment; (3) using $P(S_i)^{(t)}$, $P(A_i)^{(t)}$ and/or the previous value for the control parameter, $P(A_i \mid S_i)^{(t)}$, in the definition of the size of the adjustment; and (4) the presence or absence of an extra parameter with a value set by the testing agency to boost the adjustments. Not all possible combinations are addressed because some of them lead to conflicting results, for example, adjustments not consistent with the range of possible values for the parameters.

The first alternative to (7) is:

$$P^{(t+1)}(A_i \mid S_i) := \begin{cases} P^{(t)}(A_i \mid S_i) & \text{if } P^{(t)}(A_i) \leq r_{\max}, \\ r_{\max}/P^{(t)}(A_i) & \text{if } P^{(t)}(A_i) > r_{\max}. \end{cases} \tag{9}$$

In this adjustment rule, only negative adjustments are possible; if these adjustment lead to an exposure rate below $r_{\max}$, the control parameters are never adjusted back to 1. As a consequence, it is impossible to undo an earlier negative adjustment for an item by a positive adjustment at a later step. Hence, iterative processes in which alternate negative and positive adjustments lead to exposure rates that jump up and down may be avoided.

In addition, the adjustments are based on the exposure rates of the items, $P^{(t)}(A_i)$, instead of their probabilities of selection, $P^{(t)}(S_i)$, in two different ways: First, the size of the negative adjustment is based on $P^{(t)}(A_i)$. However, from (4),

$$\frac{r_{\max}}{P(S_i)} \leq \frac{r_{\max}}{P(A_i)}. \tag{10}$$

The adjustment for items with overexposure based on $P(S_i)$ in (9) is thus less rigorous than the same adjustment in the SH method based on $P(A_i)$. Second, from the inequality in (4) it follows that the condition under which the negative adjustment is applied is more restrictive if it is based on $P(A_i)$ than on $P(S_i)$. For the SH method it is thus possible that items with $P^{(t)}(A_i)$ already below $r_{\max}$ are nevertheless adjusted. Our evaluation of the adjustment rule in (9) relative to the SH method should thus be mixed: On one hand, this rule may need more iteration steps because its adjustments are in smaller steps, but fewer steps may suffice because it is focused better on items that need adjustment.

It seems interesting to compare the adjustment rule in (9) with the following rule:

$$P^{(t+1)}(A_i \mid S_i) := \begin{cases} P^{(t)}(A_i \mid S_i) & \text{if } P^{(t)}(S_i) \leq r_{\max}, \\ r_{\max}/P^{(t)}(S_i) & \text{if } P^{(t)}(S_i) > r_{\max}. \end{cases} \tag{11}$$

This rule may need fewer steps because its adjustments are more rigorous but is less well focused on items that need adjustment. Observe that this rule is the original SH rule modified to have only negative adjustments.

The SH rule and the two previous alternatives are based entirely either on $P(S_i)$ or $P(A_i)$. Nevertheless, as shown in the following alternative rule, it is possible to combine more rigorous negative adjustments with application of these adjustments only to items that have exposure rates $P^{(t)}(A_i) > r_{\max}$:

$$P^{(t+1)}(A_i \mid S_i) := \begin{cases} P^{(t)}(A_i \mid S_i) & \text{if } P^{(t)}(A_i) \leq r_{\max}, \\ r_{\max}/P^{(t)}(S_i) - \gamma & \text{if } P^{(t)}(A_i) > r_{\max}, \end{cases} \tag{12}$$

where $0 \leq \gamma < r_{\max}/P^{(t)}(A_i)$ is a parameter the experimenter can use to boosts the size of the adjustment.

A subtle variation on (12) is possible that allows us to study the effects of the presence of positive adjustments in the SH rule. Suppose we have a minimum exposure rate, $r_{\min}$, below which we do not want the exposure rates of items with a tendency of overexposure to settle. For an appropriate choice of constants δ and ε the following adjustment rule may realize this goal:

$$P^{(t+1)}(A_i \mid S_i) := \begin{cases} \min\{P^{(t)}(A_i \mid S_i) + \varepsilon, 1\} & \text{if } P^{(t)}(A_i) < r_{\min}, \\ P^{(t)}(A_i \mid S_i) & \text{if } r_{\min} \leq P^{(t)}(A_i) \leq r_{\max}, \\ r_{\max}/P^{(t)}(S_i) - \delta & \text{if } P^{(t)}(A_i) > r_{\max}, \end{cases} \tag{13}$$

with $0 \leq \varepsilon < \delta < r_{\max}/P^{(t)}(S_i)$. The difference is the additional adjustment for items with $P^{(t)}(A_i) < r_{\min}$. For exposure rates below this bound, if the control parameter at the previous step was below 1-ε, the control parameter is adjusted positively by a quantity ε. If at the previous step the control parameter was larger, it is set at 1. The choice of ε is critical in that this parameter has to discriminate between items with a tendency to be hardly exposed, that is, items with $P^{(t)}(A_i \mid S_i)$ close to 1, and items that had earlier adjustments of their control parameters because of a tendency to overexposure.

Finally, an entirely different type of adjustment rule is suggested. The rule is based on the earlier observation that in CAT typically a small subset of the items are popular

because they have maximum information over an interval of θ values whereas the others are seldom chosen. As a consequence, the items for which the exposure rates have to be adjusted have a probability of selection, $P(S_i)$, close to or equal to 1. However, from (5) it follows that

$$P(S_i) \rightarrow 1 \tag{14}$$

implies

$$P(A_i) \rightarrow P(A_i \mid S_i). \tag{15}$$

An possible adjustment rule therefore seems

$$P^{(t+1)}(A_i \mid S_i) := \begin{cases} P^{(t)}(A_i \mid S_i) & \text{if } P^{(t)}(A_i) \leq r_{max}, \\ P^{(t)}(A_i \mid S_i) - P^{(t)}(A_i) + r_{max} - \varphi & \text{if } P^{(t)}(A_i) > r_{max}, \end{cases} \tag{16}$$

where $0<\varphi < r_{max}$.

This rule is motivated by the fact that if at Step t

$$P^{(t)}(A_i) \rightarrow P^{(t)}(A_i \mid S_i) \tag{17}$$

and $P^{(t)}(A_i) > r_{max}$, it follows from (16) that

$$P^{(t+1)}(A_i) \rightarrow P^{(t+1)}(A_i \mid S_i) \rightarrow r_{max} - \varphi. \tag{18}$$

Parameter φ has been added to the rule in (16) to get the limit in (18) below r_{max}; choosing a lower value for φ will make the adjustment more rigorous. To deal with items for which the approximation in (14) implies a nonnegligible overestimation of their probability of selection, it seems sensible to impose the condition $P^{(t)}(A_i \mid S_i) - P^{(t)}(A_i) + r_{max} - \varphi \geq \mu$ on the adjustment for $P^{(t)}(A_i) > r_{max}$ in (16), where μ is a lower bound set by the testing agent. If the condition does not hold, $P^{(t+1)}(A_i \mid S_i)$ is

set equal to the lower bound:

$$P^{(t+1)}(A_i \mid S_i) := \mu. \tag{19}$$

Parameter μ serves thus as a lower bound to the projected exposures rates of items with an actual probability of selection that deviates too much from the limit in (14).

Simulation Studies

The goal of these simulation studies was to examine the behavior of the alternative adjustment rules in (9) and (11)-(16) relative to the original SH method. The item pool was one drawn from the Law School Admission Test (LSAT) with 397 items fitting the IRT model in (1). The adaptive test was a 30-item test from this pool. The responses on the items were simulated for examinees with θ randomly sampled from $N(0, 1)$. To reduce the number of computer simulations, exposure control was only for the entire population of examinees and not conditional on θ . The intention was to get an impression of the speed of convergence and the efficacy of the methods, and there is no reason to expect that conditional control (Stocking & Lewis, 1998, 2000) would lead to a different impression.

The ability of the examinee was estimated using the EAP estimator with a uniform prior on [-5,5]. The initial value of the ability estimator was $\widehat{\theta}_0 = 0$. The first 5 items were randomly selected from the 25 items with the value for the item difficulty parameters b_i closest to $\widehat{\theta}_0$. These items thus had a guaranteed exposure rate equal to .20. During the simulations the exposure control parameters for these items were not adjusted. For each method, the adjustment process was stopped after 25 iterations. However, because for all conditions the pattern of results did not show any meaningful change after 15 iterations, only graphs with the results for the first 15 iterations are presented. The number of examinees sampled in the CAT simulation at each step was equal to 4,000. This number was chosen to be extremely large to minimize the effects of sampling error in the estimates of $P(A_i)$ and $P(S_i)$ on the behavior of the adjustment rules. For all exposure control methods, the target for the exposure rates was $r_{max} = .2$.

First Study

In the first study, the behavior of the SH method and the alternative adjustment rules in (9) and (11)-(16) was compared. For the rules in (12), (13), and (16) we used $\gamma = .15$, $\delta = .15$, r_{min}=.10, $\varepsilon = .10$, $\varphi = .10$, and $\mu = .10$. The value for r_{min} was larger than the bound in (A3), but the rule in (13) imposes this value only on the exposure rates of the (small) subset of items in the pool that have a tendency to overexposure.

[Figure 1 about here]

Figures 1 shows the number of items that violated the target value, r_{max}, the maximum exposure rate, and the average exposure rate for the items that violated the target as a function of the iteration steps. For each of these criteria, the adjustment rules in (12) and (16) were superior, with the number of violations for the former converging somewhat faster than for the former but at the price of having slightly larger maximum exposure rates. Both the SH rule and the rule in (11) produced inferior results for the number of violations, but these rules had maximum exposure rates and average rates for the violators that tended to be close to those for the rules in (12) and (16). For the adjustment rules in (9) and (13) the behavior was opposite; they tended to produce numbers of violations that were second best but had the worst maximum exposure rates and average rates for the violators. Also, the behavior of these two rules was least smooth for all three criteria. In fact, both rules showed a strong saw-tooth pattern in their maximum exposure rates and average exposure rates for the violators across the iteration steps. For the rule in (9), this pattern is believed to be the result from its positive adjustments for items with $P^{(t)}(A_i) < r_{min}$.

Second Study

The fact that the adjustment rules in (12) and (16) produced superior results suggested a second study in which the size of their parameters γ and φ was systematically varied. The values chosen for both γ and φ were .05, .10, .15, .20, and .,25.

[Figure 2 about here]

The results for the rule in (12) are given in Figure 2. For each of the three criteria, the results were identically ordered in the value of γ at nearly every iteration step. That is, the higher the value of γ, the better the results. For γ=.25, the adjustment rule produced admissible exposure rates for all items at the eight iteration step, but was already negligibly close to this result at the fifth step. This result is remarkable when compared with those for the SH method, which never produced fewer than 50 violators and had a lowest maximum exposure rate of .246 and average violation of .216 across all 25 iteration steps.

[Figures 3 about here]

Figure 3 shows the results for the rule in (16) as a function of φ. Generally, the results in Figure 2 and 3 are close to each other, particularly for the number of violations and average violation. The most important differences between these two rules are for the number of violations: The rate of convergence for this criteria appeared to be much more sensitive to the adjustment parameters γ for the rule in (12) than to the parameter φ for the rule in (16). For larger values of γ in (12), the number of violations became negligibly low for a smaller number of iteration steps than for the larger values of φ in (16). However, for smaller values of γ, the result was reversed. In fact, for $\gamma = .05$ the number of violations for (12) remained much too high across all iteration steps.

Discussion

The analysis of the SH method in this paper showed why the method may behave erratically in some applications and has difficulty reaching admissibility. The results from the simulation studies presented in this paper illustrate the consequences of this behavior.

The simulations studies also showed that the adjustment rules in (12) and (16) are practical alternatives to the SH method. For adjustment parameter γ=.25, the rule in (12) seems to be capable to produce admissible exposure rates for all items in the pool in 5-8 iteration steps. The rule in (16) was somewhat slower, but its results seem to be much less

sensitive to the choice of value for adjustment parameter φ. Both features are attractive. In a practical application, the choice between these two methods should thus depend on such factors as the amount of time available and the possibility to use previous experience with comparable item pools to set parameter values.

The combination of features of the two rules that may have made them superior is: (1) negative adjustments of the control parameters only for items with an exposure rate larger than the target value r_{max}; (2) more rigorous negative adjustments due to the introduction of an extra adjustment parameter; and (3) no adjustments for items with exposure rates already below the target value r_{max}. The SH method misses these three features. Of course, additional studies are needed to generalize the results to other item pools and CAT algorithms. In particular, it deserves further study to find out if the presence of content constraint on item selection in CAT would force us to revise the conclusion.

Though the adjustment rules in (12) and (16) seem to have the potential of a substantial reduction of the costs involved in the preparations of a new item pool for a CAT program, it is still desirable to look for less costly methods of item-exposure. One alternative approach is a-stratified adaptive testing (Chang & Ying, 1999), in which during the test item-selection is sequentially constrained to strata in the item pool with increasing values for the item discrimination parameter in (1), a_i. For a technique to implement these constraints on the item selection along with large numbers of other constraints, for example, on the content on the test, see van der Linden and Chang (in press). Another approach could be based on the suggestion in Wainer (2000, pp. 293-294) to define an index of test quality and retire items from the pool immediately after they have been administered once. If the index falls below a minimally acceptable level, the items are returned to the pool for one possible additional administration. Finally, an alternative approach is developed in van der Linden and Veldkamp (2002) who impose probabilistic item eligibility constraints on the selection of items for the CAT that are implemented through a shadow test approach (van der Linden, 2000). Like the Sympson-Hetter method, the decisions to impose these constraints on the items are made using a probability experiment. However, unlike this method, the probabilities in this experiment

do not need adjustment of any control parameters but are set adaptively during operational use of the item pool.

Appendix: Some Properties of Item Exposure Rates and SH Exposure Control

This appendix summarizes some of the properties of item-exposure rates and the way they are controlled in the SH method. Some of these properties are immediately obvious, but have not yet been noticed in the literature. The properties are summarized in the following set of statements

Property 1. *For any CAT algorithm, item pool, and ability distribution, it holds that*

$$\sum_{i=1}^{I} P(A_i) = n, \tag{A1}$$

where n is the (fixed) number of items in the CAT.

This property follows from the fact that each examinee encounters n items. Because n is fixed, the average encounter rate is nI^{-1}. The equality in (A1) follows because the average exposure rate of the items is equal to the average encounter rate. If n is random, the right-hand side of (A1) should be replaced by its expectation. A random test length occurs if the CAT stops as soon as a fixed level of accuracy for the ability estimator is realized. Most CAT programs have a fixed test length though, and the current paper addresses this case.

Property 2. *To obtain admissibility it is necessary that*

$$r_{\max} \geq nI^{-1}. \tag{A2}$$

From (A1) it follows that the only possible common value of $P(A_i)$ for all items is nI^{-1}. This value is a lower bound on $r_{\max}$ because it can only be realized if the exposure control parameters can be manipulated to have the algorithm sample items from the pool with equal marginal probabilities of selection.

Property 3. *To impose a minimum exposure rate, $r_{\min}$, on all items in the pool, it is necessary that*

$$r_{\min} \leq nI^{-1}. \tag{A3}$$

The fact that $nI^{-1}$is the only possible common value of $P(A_i)$ for all items implies also the upper bound on r_{min} in (A3).

Property 4. *The sum of the exposure rates remains constant across SH adjustments, that is*

$$\sum_{i=1}^{I} P^{(t+1)}(A_i) = \sum_{i=1}^{I} P^{(t)}(A_i), \text{ for } t = 1,.2, \ldots \quad \text{(A4)}$$

This property is true because the equality in (A1) holds within each iteration step.

The following two properties allow us to assess the possible effects of the SH adjustments in an earlier iteration step on the item exposure rates in a later step.

Property 5 *The effects of the changes in the exposure control parameters at Step t on the probabilities of selection of the items at Step $t+1$ satisfy the following constraint:*

$$\sum_{i=1}^{I} P^{(t+1)}(A_i \mid S_i)P^{(t+1)}(S_i) = \sum_{i=1}^{I} P^{(t)}(A_i \mid S_i)P^{(t)}(S_i). \quad \text{(A5)}$$

The constraint follows from (A4) and (5). All factors in this constraint are known fixed constants except the probabilities $P^{(t+1)}(S_i)$. The simulation at Step $t + 1$ is conducted to estimate these probabilities.

The final property suggests a hidden assumption on which the SH method seems to rely:

Property 6. *The adjustments in the SH method follow from the assumption that*

$$P^{(t+1)}(A_i) = P^{(t+1)}(A_i \mid S_i)P^{(t)}(S_i), \text{ for } i = 1, \ldots, I \text{ and } t = 1,.2, \ldots \quad \text{(A6)}$$

This result is immediately clear if we substitute the adjustments in (7) for the cases of $P^{(t)}(S_i) > r_{max}$ and $P^{(t)}(S_i) \leq r_{max}$ into (A6). These substitutions yield the desired results $P^{(t+1)}(A_i) = r_{max}$ and $P^{(t+1)}(A_i) \leq r_{max}$, respectively. The assumption in (A6) would hold if the equality in (5) were valid between probabilities in consecutive steps. However, (5) only holds within steps. If it held between steps, it would follow from (A6) that the SH method always reached admissibility in one step.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Chang, H., & Ying, Z. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211-222.

Chang, S.-W., & Harris, D. J. (2002, April). *Redeveloping the exposure control parameters of CAT items when a pool is modified.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Hetter, R. R., & Sympson, J. B. (1997). Item-exposure in CAT-ASVAB. IN W. A. Sands, J. R., Waters & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141-144). Washington, DC: American Psychological Association.

Mislevy, R. J., & Chang, H. (2000). Does adaptive testing violate local independence? *Psychometrika, 65*, 149-156.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57-75.

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163-182). Boston: Kluwer.

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27-52). Boston: Kluwer.

van der Linden, W. J., Chang, H. (in press). Implementing content constraints in alpha-stratified adaptive using a shadow test approach. *Applied Psychological Measure-*

ment.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1-25). Boston: Kluwer.

van der Linden, W. J., & Veldkamp, B. P. (2002). *Constraining item exposure in computerized adaptive testing with shadow tests*. Manuscript to be submitted for publication.

Wainer, H. (2000a). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Erlbaum.

Wainer, H. (2000b). Rescuing computerized testing by breaking Zipf's law. *Journal of Educational and Behavioral Statistics, 25*, 203-224.

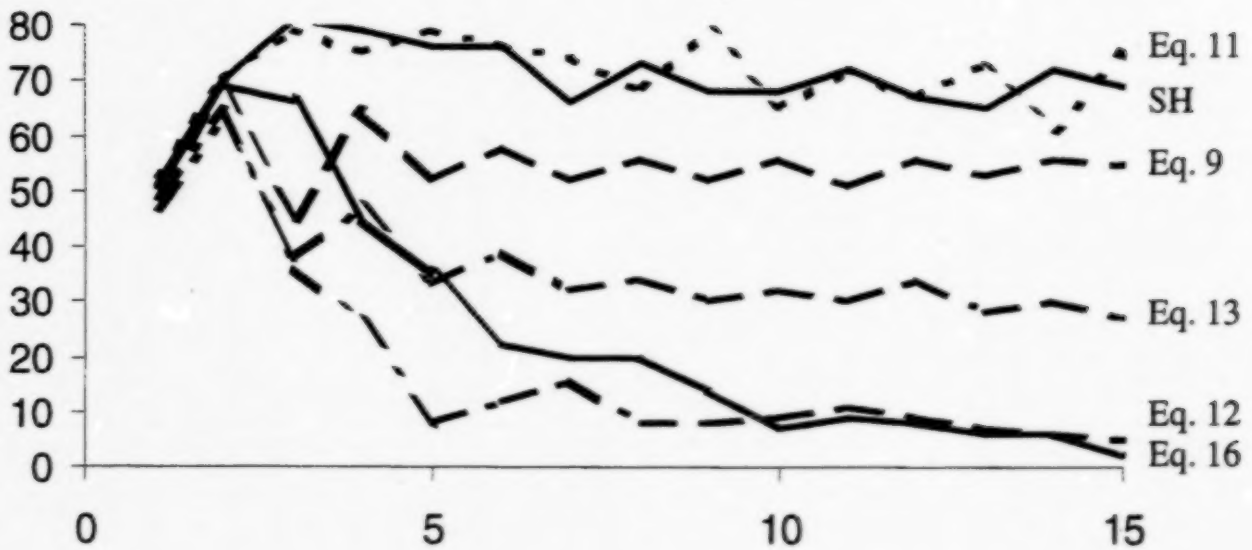
Figure Captions

Figure 1. Number of violators of target exposure rate (panel a), maximum exposure rate (panel b) and average exposure rate (panel c) as a function of the iteration steps for all six exposure control methods .

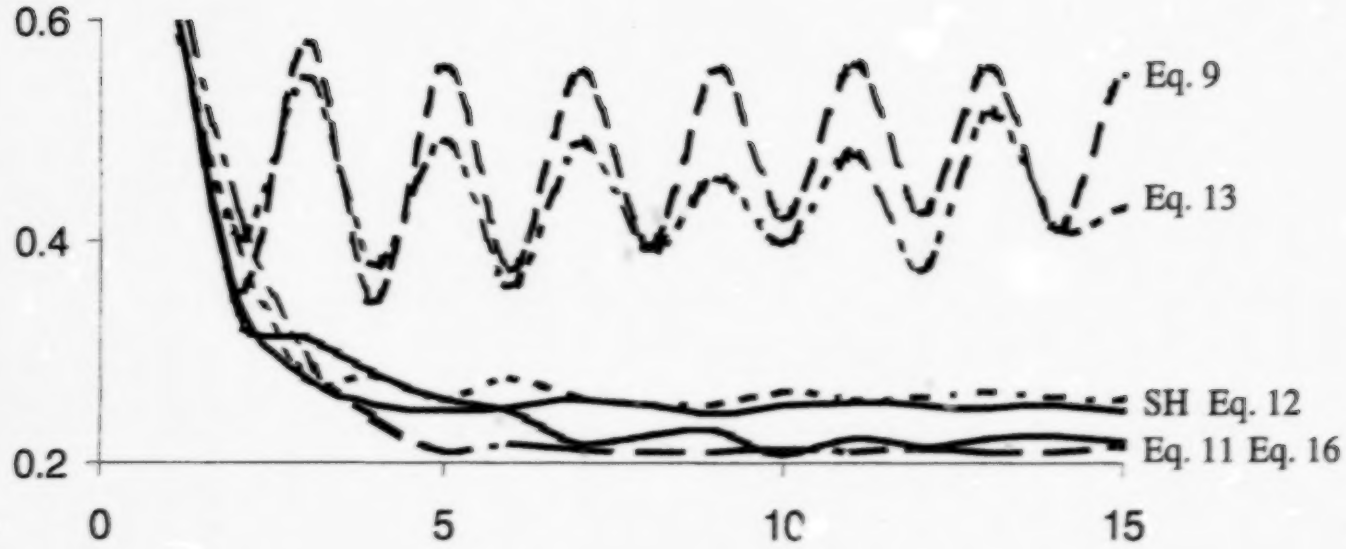
Figure 2. Number of violators of target exposure rate (panel a), maximum exposure rate (panel b) and average exposure rate (panel c) as a function of the iteration steps for adjustment parameter γ=.05, .10, .15, .20 and .25 in the exposure control method in Equation 12.

Figure 3. Number of violators of target exposure rate (panel a), maximum exposure rate (panel b) and average exposure rate (panel c) as a function of the iteration steps for adjustment parameter φ=.05, .10, .15, .20 and .25 in the exposure control method in Equation 16.

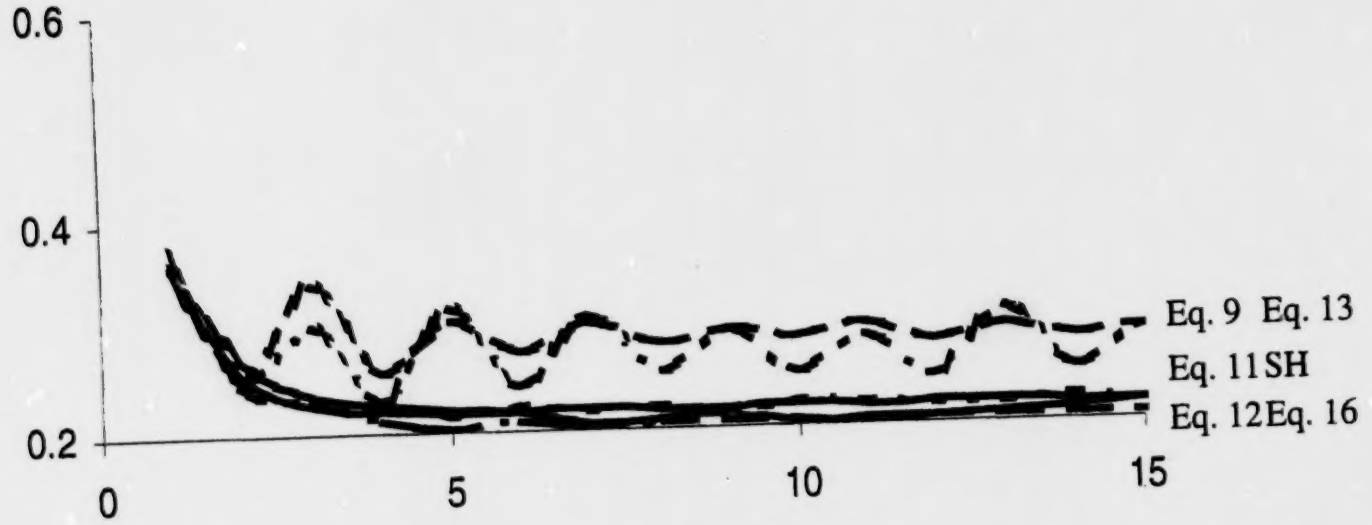
(a)



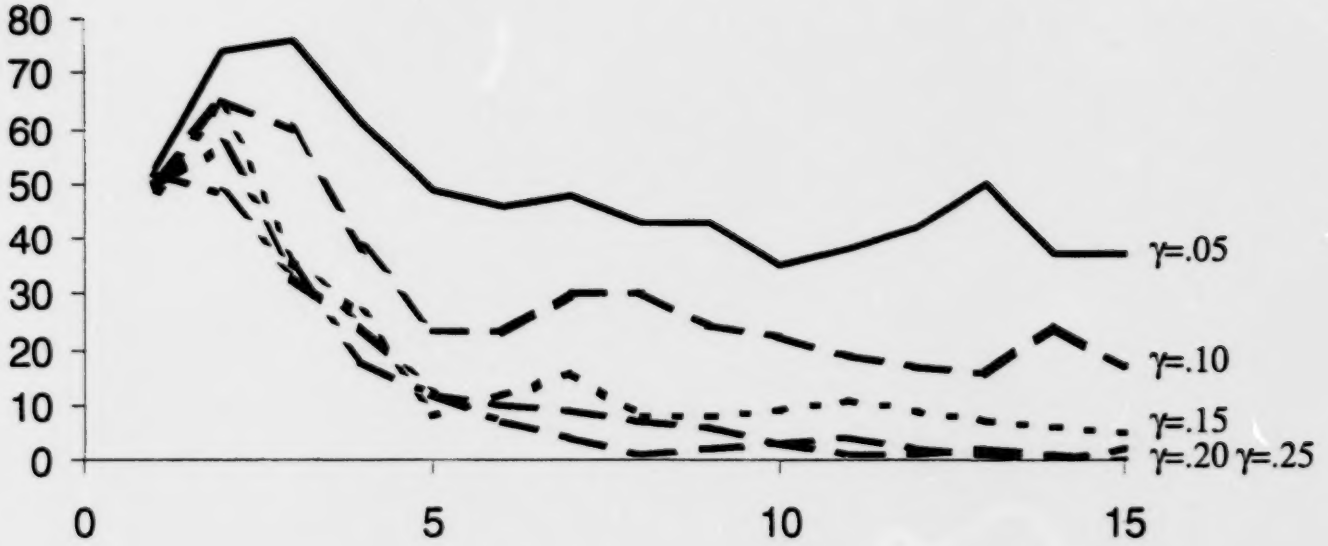
(b)



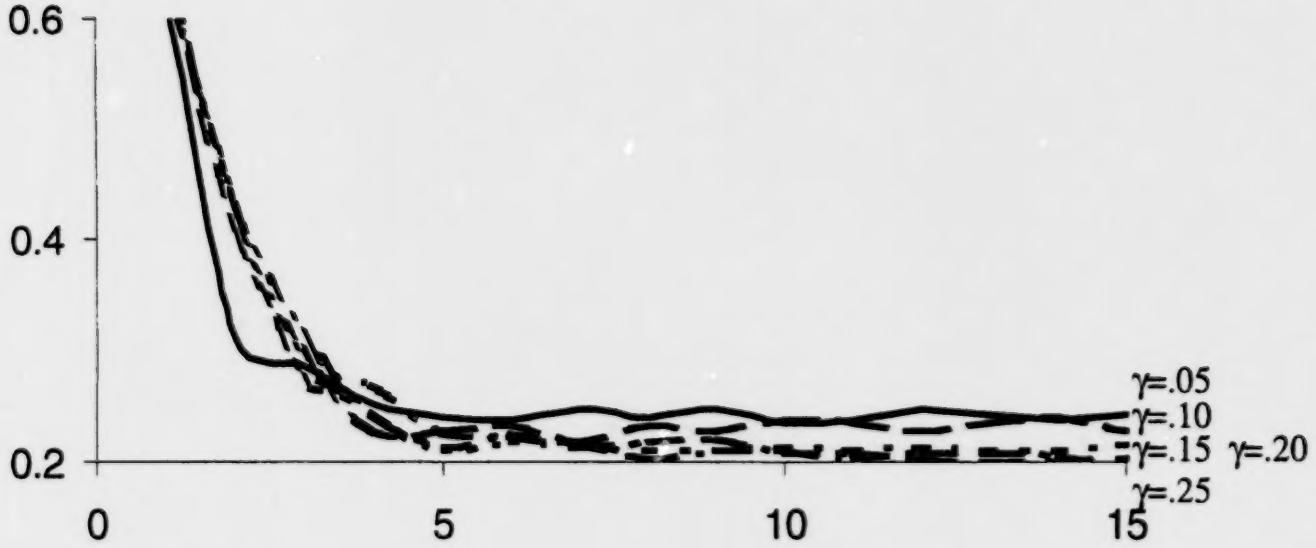
(c)



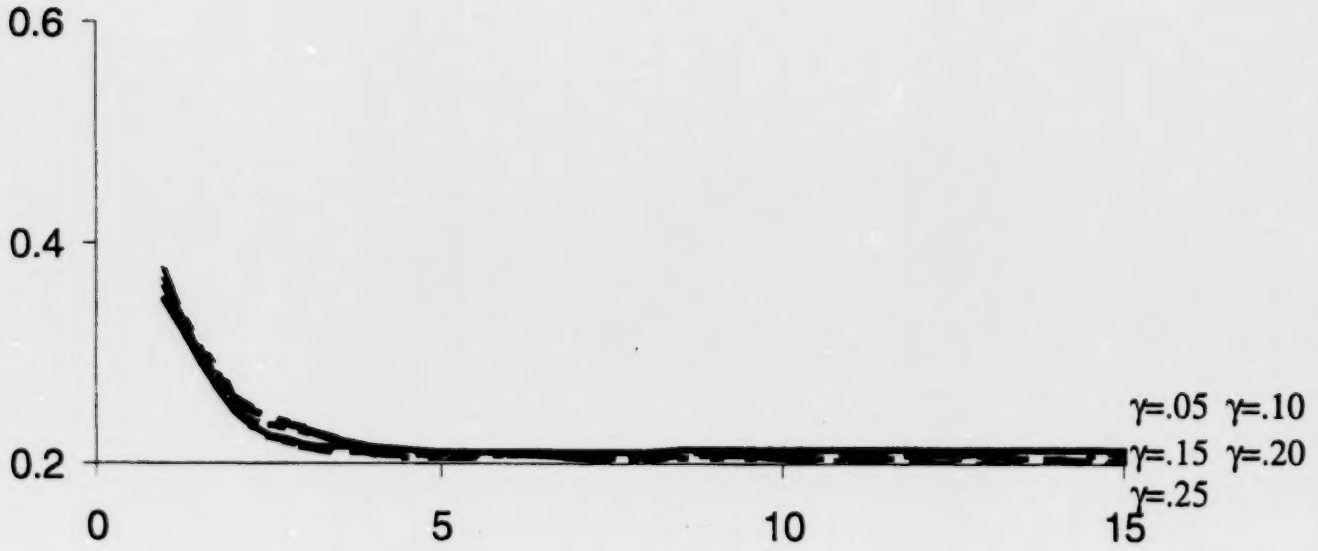
(a)



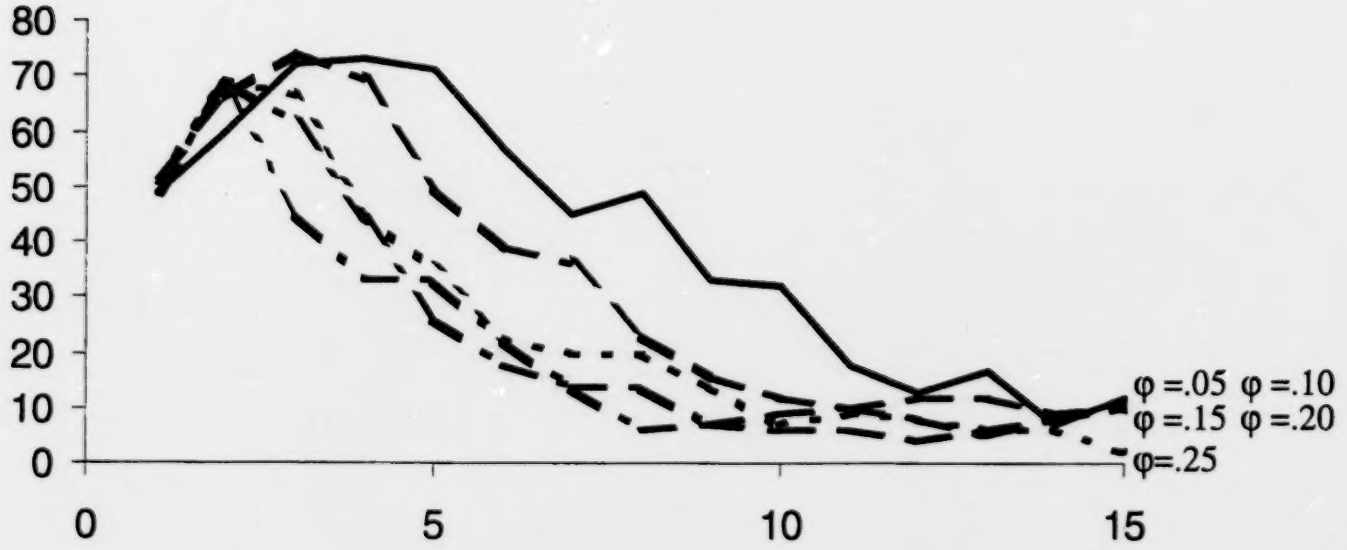
(b)



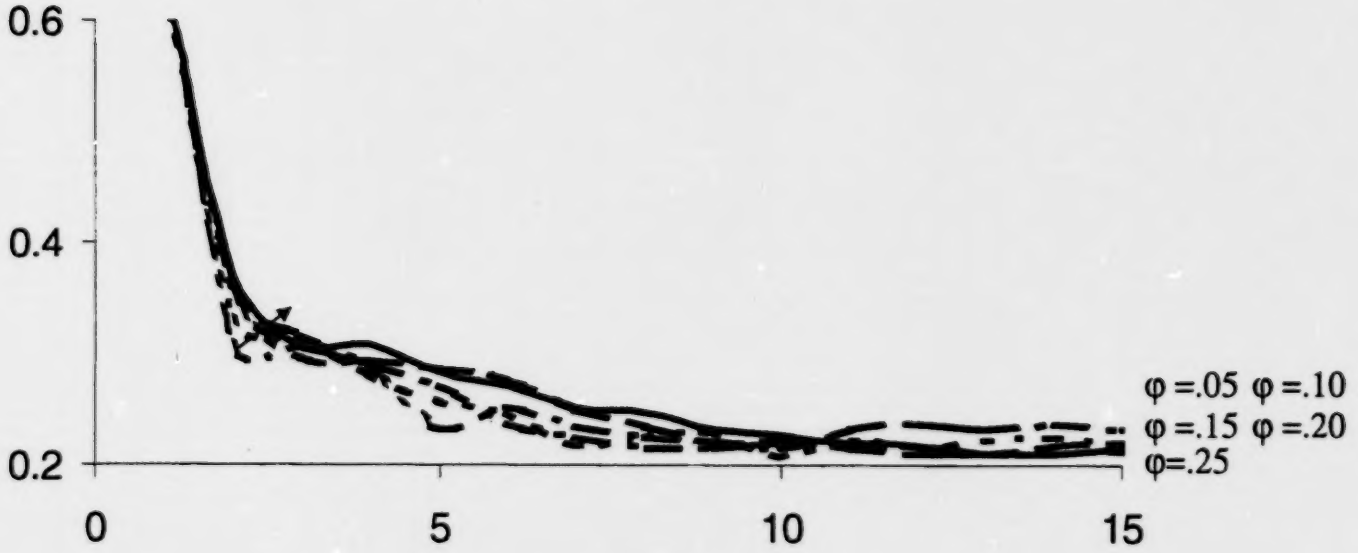
(c)



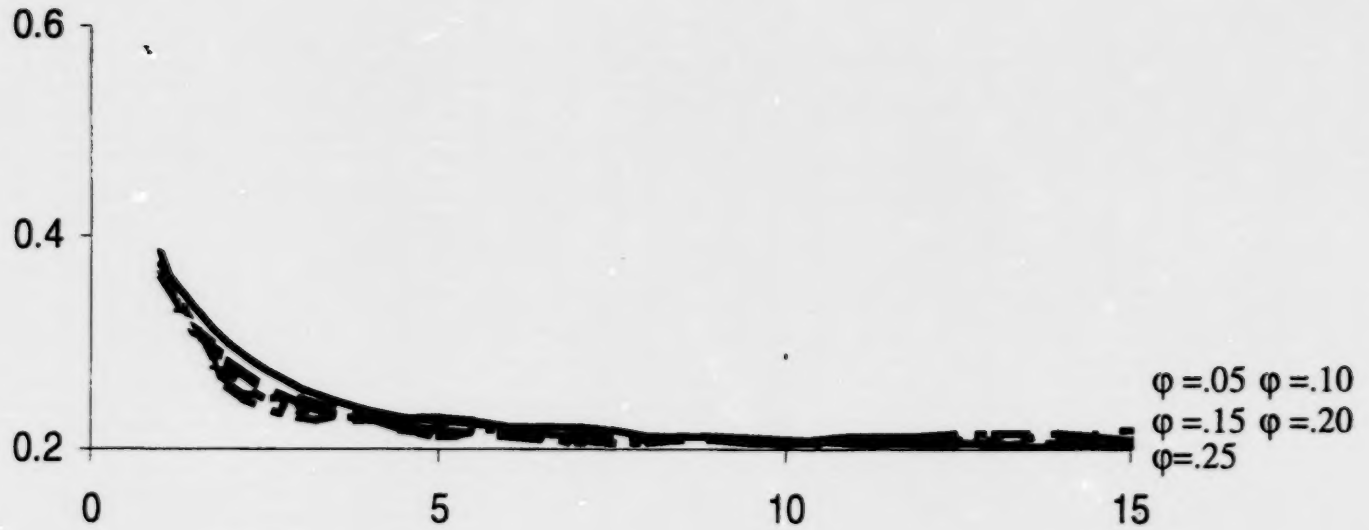
(a)



(b)



(c)



Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-02-02 W.J. van der Linden, *Some Alternatives to Sympson-Hetter Item-Exposure Control in Computerized Adaptive Testing*

RR-02-01 W.J. van der Linden, H.J. Vos, & L. Chang, *Detecting Intrajudge Inconsistency in Standard Setting using Test Items with a Selected-Response Format*

RR-01-11 C.A.W. Glas & W.J. van der Linden, *Modeling Variability in Item Parameters in Item Response Models*

RR-01-10 C.A.W. Glas & W.J. van der Linden, *Computerized Adaptive Testing with Item Clones*

RR-01-09 C.A.W. Glas & R.R. Meijer, *A Bayesian Approach to Person Fit Analysis in Item Response Theory Models*

RR-01-08 W.J. van der Linden, *Computerized Test Construction*

RR-01-07 R.R. Meijer & L.S. Sotaridona, *Two New Statistics to Detect Answer Copying*

RR-01-06 R.R. Meijer & L.S. Sotaridona, *Statistical Properties of the K-index for Detecting Answer Copying*

RR-01-05 C.A.W. Glas, I. Hendrawan & R.R. Meijer, *The Effect of Person Misfit on Classification Decisions*

RR-01-04 R. Ben-Yashar, S. Nitzan & H.J. Vos, *Optimal Cutoff Points in Single and Multiple Tests for Psychological and Educational Decision Making*

RR-01-03 R.R. Meijer, *Outlier Detection in High-Stakes Certification Testing*

RR-01-02 R.R. Meijer, *Diagnosing Item Score Patterns using IRT Based Person-Fit Statistics*

RR-01-01 H. Chang & W.J. van der Linden, *Implementing Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow Test Approach*

RR-00-11 B.P. Veldkamp & W.J. van der Linden, *Multidimensional Adaptive Testing with Constraints on Test Content*

RR-00-10 W.J. van der Linden, *A Test-Theoretic Approach to Observed-Score Equating*

RR-00-09 W.J. van der Linden & E.M.L.A. van Krimpen-Stoop, *Using Response Times to Detect Aberrant Responses in Computerized Adaptive Testing*

RR-00-08 L. Chang & W.J. van der Linden & H.J. Vos, *A New Test-Centered Standard-Setting Method Based on Interdependent Evaluation of Item Alternatives*

RR-00-07 W.J. van der linden, *Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing*

RR-00-06 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory*

RR-00-05 B.P. Veldkamp, *Modifications of the Branch-and-Bound Algorithm for Application in Constrained Adaptive Testing*

RR-00-04 B.P. Veldkamp, *Constrained Multidimensional Test Assembly*

RR-00-03 J.P. Fox & C.A.W. Glas, *Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory*

RR-00-02 J.P. Fox, *Stochastic EM for Estimating the Parameters of a Multilevel IRT Model*

RR-00-01 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*

RR-99-08 W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*

RR-99-07 N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*

RR-99-06 G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*

RR-99-05 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*

RR-99-04 H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

RR-99-03 B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*

RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*

RR-99-01 R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

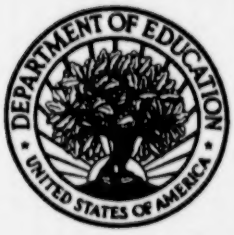
Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.

faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede

TM034737



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").